Performance of a Rasch-based method for group comparisons of longitudinal change and response shift at the item level in PRO data: a simulation study

Myriam Blanchin, Priscilla Brisson, Véronique Sébille

Methods 204 (2022) 327–339

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Performance of a Rasch-based method for group comparisons of longitudinal change and response shift at the item level in PRO data: A simulation study

Myriam Blanchin [a,*], Priscilla Brisson [a], Véronique Sébille [a,b]

[a] U1246 SPHERE "methodS in Patient-centered outcomes and HEalth ResEarch", Université de Nantes, Université de Tours, INSERM, Nantes, France
[b] Methodology and Biostatistics unit, CHU of Nantes, Nantes, France

ARTICLE INFO

Keywords:
Response shift
Patient-reported outcomes
Rasch models
Longitudinal data
Item level

ABSTRACT

The growing interest in patient perception and experience in healthcare has led to an increase in the use of patient-reported outcomes (PRO) data. However, chronically ill patients may regularly adapt to their disease and, as a consequence, might change their perception of the PRO being measured. This phenomenon named response shift (RS) may occur differently depending on clinical and individual characteristics.

The RespOnse Shift ALgorithm at the Item level (ROSALI), a method for RS analysis at the item level based on Rasch models, has recently been extended to explore heterogeneity of item-level RS between two groups of patients. The performances of ROSALI in terms of RS detection at the item level and biases of estimated differences in latent variable means were assessed.

A simulation study was performed to investigate four scenarios: no RS, RS in only one group, RS affecting both groups either in a similar or a different way. Performances of ROSALI were assessed using rates of false detection of RS when no RS was simulated and a set of criteria (presence of RS, correct identification of items and groups affected by RS) when RS was simulated.

Rates of false detection of RS were low indicating that ROSALI satisfactorily prevents from mistakenly inferring RS. ROSALI is able to detect RS and identify the item and group(s) affected when RS affects all response categories of an item in the same way. The performances of ROSALI depend mainly on the sample size and the degree of heterogeneity of item-level RS.

1. Introduction

The growing interest in patient perception and experience in healthcare has led to an increase in the use of patient-reported outcomes (PRO) data to incorporate the patient perspective into clinical care, in clinical trials and in healthcare policy [1–4]. PRO instruments can be used to measure unobservable constructs such as health-related quality of life (HRQoL), fatigue or anxiety. Such constructs are often referred to as "latent traits" and they are measured via self-reported questionnaires in which items are often grouped within several domains (physical, emotional, social…). PRO data come directly from patients without involving the perspective of anyone else and aim to reflect patient's own experience of illness.

Longitudinal PRO data are of value for the analysis and interpretation of PRO change over time in epidemiological or clinical research studies following a specific health event, e.g. diagnosis, treatment initiation. However, PRO data remain difficult to analyze and interpret. Indeed, in the context of chronic disease, for instance, patients may have to regularly adapt to their illness. For example, patients might experience levels of acute pain that they had never experience before their surgery. As a consequence, they might change their perception of the construct to be measured (e.g. chronic pain) and of the items reflecting it. Hence, in case of a change in perception of the PRO, longitudinal PRO data may not be comparable over time (e.g. scores of chronic pain before and after surgery) due to lack of measurement invariance, also called response shift (RS) whose definition has been recently updated by Vanier et al. [5]. RS is considered to be a "special case of violation of the principle of conditional independence when observed change is not fully explained by target change" (i.e. change in the construct of interest) as a result of "a change in meaning in self-evaluation of a target construct" [5]. Violation of the principle of conditional independence may be due

* Corresponding author.
E-mail address: myriam.blanchin@univ-nantes.fr (M. Blanchin).

https://doi.org/10.1016/j.ymeth.2022.01.002
Received 9 September 2021; Received in revised form 20 December 2021; Accepted 3 January 2022
Available online 5 January 2022

to how RS manifests itself through recalibration (a change in one's internal standards), reprioritization (a change in one's values), and reconceptualization related to one's redefinition of the target construct [6].

In case of RS, the estimation of longitudinal change in PRO data can be biased if it is not accounted for and it may threaten the interpretation of change and the assessment of possible intervention effects [7]. Influence of RS on PRO evaluation may lead to suboptimal medical decisions on both the individual patient and health policy levels [8]. At the patient/clinician level, shared decision making may not be fully informed if it is based on previous published PRO studies not accounting for RS. At healthcare policy level, guidelines on treatment preference can be obfuscated by RS. However, RS could also be a result of positive adaptations to health challenges and can also reveal maladaptive disorders that are worth detecting [9,10]. Whatever the adopted viewpoint, it seems important to detect and quantify RS in a reliable manner.

Most statistical methods proposed for RS detection are performed at the domain level [11,12]. These methods consist in the analysis of sum scores that summarize the information of item responses into one value. In particular, patients with different response profiles can have the same sum score. As domain-level RS analyses cannot distinguish which items are specifically affected by RS, Schwartz et al. [13] suggested that analyzing RS at the item level could provide additional information for the interpretation of RS effects [13]. Different methods have been proposed for RS detection at the item level based on different latent variable models (Structural Equation Modelling, Item Response Theory and Rasch Measurement Theory) where the latent variable represents the unobservable PRO of interest (e.g. HRQoL). They were compared recently [14] in a simulation study and the method based on Rasch Measurement Theory (RMT) models, called the RespOnse Shift ALgorithm at the Item level (ROSALI) showed better performances compared to other methods for detecting and accounting for recalibration (RC) in the measurement of PRO change. Rates of incorrect detection of RC of ROSALI ranged from 0.6% to 2.6% and rates of correct detection of RC ranged from 83.2% to 100% depending on the questionnaire length and the number of response categories. However, in this former version of ROSALI it was assumed that the majority of patients experiences RS the same way. This restrictive assumption of the homogeneity of RS within a sample is often made when using most RS detection methods whether they be at the item level or at the domain level [12]. However, change in interpretation of items may be influenced by cultural, or personality differences, as well as life circumstances, and/or because of different health experiences or events. A study assessing self-reported depression before and after several treatments (e.g. cognitive behavioral therapy, antidepressant) evidenced RS leading to overestimation of depressive symptomatology after the treatment period [15]. Heterogeneity in RS effects was suggested as RS seemed to be higher for patients in the psychotherapy groups probably because they have received more psychoeducation than patients receiving only medication. To date, the effect of measured or unmeasured covariates on RS in longitudinal PRO data has been investigated using Structural Equation Modelling by incorporating covariates in the analysis [16], performing stratified analysis [9] or using a combination of Mixed Models and Growth Mixture Models and Structural Equation Modelling [10]. Nevertheless, all these studies were performed at the domain level. Hence, new developments are needed to account for RS heterogeneity at the item level. ROSALI has thus been extended to explore the heterogeneity of item-level RS between groups in studies comparing two groups of patients [17]. For example, patients from two different treatment groups might experience their illness in a different way and RS may occur differently in each group or even occur in only one treatment group.

In studies comparing two groups of patients, the perception of items might also be different from one group to another at a specific time, a phenomenon referred as differential item functioning (DIF). DIF analyses are frequently used in cross-sectional studies to assess if some items display DIF according to some covariates (e.g. cancer sites, age, gender) [18,19]. DIF can also bias the estimation of the difference in PRO between groups [20]. Hence, in longitudinal studies comparing two groups of patients between time 1 and time 2, DIF should also be considered along with RS. Therefore, ROSALI was not only extended to explore different RS between groups but also to assess whether some items function differently between groups at time 1. To date, the latest version of ROSALI enables to detect and adjust for DIF and RS in the estimation of PRO change to ensure valid comparisons between groups and over time. These major changes of ROSALI were described elsewhere [17] alongside with an illustrative example of interpretation of the results of the algorithm.

Performances of methods for RS detection have rarely been assessed and conducting simulation studies have been recommended to fill this gap [11,12]. Although the performances of ROSALI without a covariate were satisfactory in a previous simulation study [14], the effects of the major changes made to extend ROSALI for longitudinal studies with two groups need to be assessed. The aim of this article is thus to assess the performances of ROSALI in terms of RC detection at the item level in the context of longitudinal studies designed for the comparison of two groups of patients using a simulation study.

2. Methods

ROSALI and the simulated datasets for the simulation study are based on RMT models. Models from RMT assume a non-linear link between observed item responses and the unobservable latent variable that represents the PRO of interest (e.g. HRQoL).

2.1. Generation of data

Simulated datasets were composed of answers of N patients in each of 2 groups ($g = 0$ *or* 1) responding at two time ($t = 1$ *or* 2) to J polytomous items with M response categories. A partial credit model (PCM) from RMT was used to model patients' responses to polytomous items [21,22] as a function of the latent variable ($\theta_i^{(t)}$) of each patient i at time t and of the threshold parameters ($\delta_{jpg}^{(t)}$) of item j for response category p in group g at time t. Each item had the same number of response categories (M), 0 is considered as the least favorable response (negative response) with respect to the latent variable (e.g. poor HRQoL), the other responses were ordered going from 1 to $(M-1)$ (i.e. $M-1$ possible positive responses). Fig. 1 presents the category probability curves of an item j at two different times. For each response category, a curve represents the probability for a patient to endorse this response category as a function of his/her level on the latent variable at a given time. Threshold parameters are operationalized as the intersections of the probability curves of two adjacent response categories; it represents the latent variable level for which a patient has the same probability of choosing one or the other adjacent response category. The higher the threshold parameter, the higher the level of the latent variable must be to endorse this response category.

The probability for patient i, belonging to group g, to answer x to item j at time t is given by:

$$P\left(X_{ij}^{(t)} = x | \theta_i^{(t)}, \beta, g_i, \beta_{inter}, t_2, \delta_{j1g}^{(t)}, \cdots, \delta_{j(\mathrm{M}-1)g}^{(t)}\right) = \frac{\exp\left(x\left(\beta \times g_i + \beta_{inter} \times t_2 \times g_i + \theta_i^{(t)}\right) - \left(\sum_{p=1}^{x} \delta_{jpg}^{(t)}\right)\right)}{\sum_{l=0}^{(\mathrm{M}-1)} \exp\left(l\left(\beta \times g_i + \beta_{inter} \times t_2 \times g_i + \theta_i^{(t)}\right) - \left(\sum_{p=1}^{l} \delta_{jpg}^{(t)}\right)\right)}$$

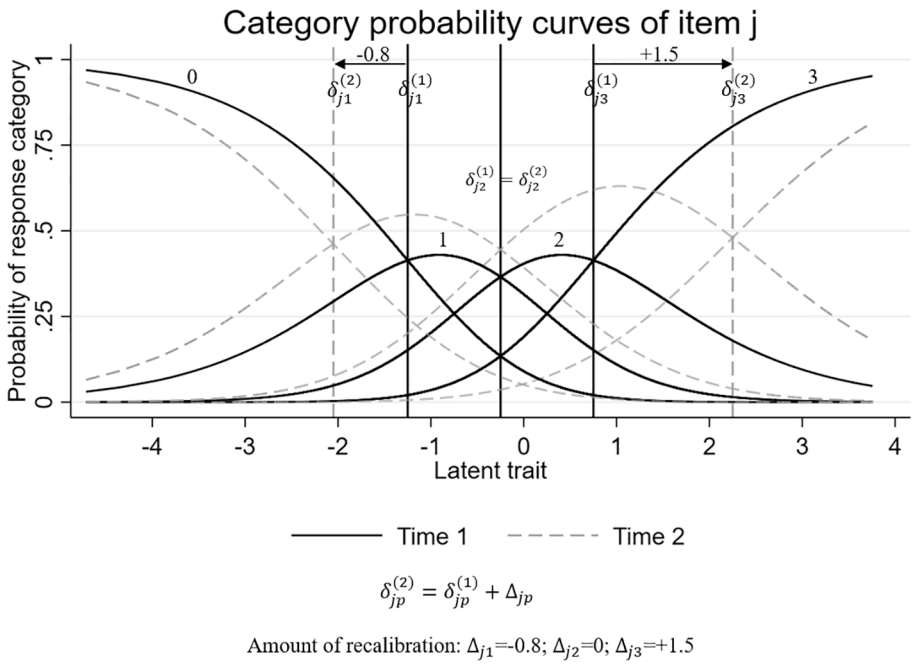

Fig. 1. Example of category probability curves of an item j for 2 times of measurement. Example for one item with 4 response categories: probability curves obtained with a Partial Credit Model (PCM). $\delta_{jp}^{(t)}$: the p[th] threshold parameter of item j at time t– defined by the equal probability to answer to two adjacent categories. As threshold parameters change over time, recalibration occurs on this item j. Δ_{jp} : amount of recalibration for the p[th] threshold parameter of item j.

With $\begin{bmatrix}\Theta^{(1)}\\ \Theta^{(2)}\end{bmatrix} \sim N\left(\begin{bmatrix}\mu^{(1)}\\ \mu^{(2)}\end{bmatrix}; \Sigma\right)$

The latent variable Θ is a random variable assumed to be normally distributed with mean vector $\begin{bmatrix}\mu^{(1)}\\ \mu^{(2)}\end{bmatrix}$ where $\mu^{(1)}$ and $\mu^{(2)}$ are the means of the latent variable at time 1 and time 2 respectively, and the covariance matrix is $\Sigma = \begin{bmatrix}\sigma_1^2 & \sigma_{1,2}\\ \sigma_{2,1} & \sigma_2^2\end{bmatrix}$. $\theta_i^{(t)}$ is the realization of Θ for patient *i* at time t.

We have : $X_{ij}^{(t)}$ the response of patient *i* to item *j* at time *t*; g_i the group indicator variable for patient *i* (0 if patient *i* is in group 0, 1 otherwise); t_2 the time indicator variable ($t_2 = 1$ for time 2, 0 otherwise); β the group effect parameter; β_{inter} the time × group interaction parameter. For group 0: $\mu_0^{(1)} = \mu^{(1)}$ and $\mu_0^{(2)} = \mu^{(2)}$(time effect) and for group 1 : $\mu_1^{(1)} = \mu^{(1)} + \beta$ and $\mu_1^{(2)} = \mu^{(2)} + \beta + \beta_{inter}$. $\delta_{jpg}^{(t)}$ is the p^{th} threshold parameter for item *j* in group *g* at time *t* ($p > 0$).

The effects of sample size, size of the questionnaire and number of response categories were investigated in the simulation study. Their simulated values were based on sample sizes and questionnaire characteristics that can be encountered in studies assessing PRO. No group effect ($\beta = 0$) or a small group effect ($\beta = 0.2$) as well as no time effect ($\mu^{(2)} = 0$) or a medium time effect ($\mu^{(2)} = 0.3$) were simulated on the latent variable. Values of the simulation parameters (sample size N, distribution of the latent variable: means $\mu_g^{(t)}$, g = 0,1 and t = 1,2, and covariance matrix Σ, number of items J and response categories M) used to generate the datasets are presented in Table 1.

Table 1
Simulation parameters.

Parameters	Simulated values
Sample size (N) with 2 equal group sizes	200; 300; 500
Number of items (J)	4; 7
Number of response categories (M)	4; 7
Latent variable mean in group 0 at time 1 ($\mu_0^{(1)}$)	0
Group effect ($\beta = \mu_1^{(1)}$)	0; 0.2
Time effect ($\mu_0^{(2)}$)	0; 0.3
Time × group interaction (β_{inter})	0
Covariance matrix (Σ)	$\begin{bmatrix}1 & 0.6\\ 0.6 & 1\end{bmatrix}$

2.2. Threshold parameters

Threshold parameters were chosen in such a way that items' distribution is centered on the mean of latent variable at time 1 [14,23]. This reflects the situation where the questionnaire is suitable for a population with a normally distributed latent variable.

2.2.1. Time 1

Threshold parameters were computed as follows. The first threshold of each item (δ_{j1}^*) was initialized with the $\frac{j}{J+1}th$ percentile from a standard normal distribution. Other threshold parameters were regularly spaced so that threshold parameters of an item *j* have a range of 2 with $\delta_{jp}^* = \delta_{j1}^* + 2 \times \frac{p-1}{M-2}$ ($p \in [2; M-1]$). Threshold parameters were finally centered on the same mean as the latent variable in group 0 (0) by subtracting the mean of all threshold parameters of all items $\delta_{jp} = \delta_{jp}^* - \overline{\delta_{jp}^*}$. Values for threshold parameters used at time 1, as a function of the number of items and response categories, are presented in Table 2, (e.g. when J = M = 4, threshold parameters of item 2 were −1.25, −0.25, and 0.75).

No difference in threshold parameters of any item between groups at time 1 was simulated, that is: $\delta_{jp0}^{(1)} = \delta_{jp1}^{(1)} \forall p, j$. Hence, no DIF was simulated at time 1.

2.2.2. Time 2

RC, a change in internal standards of measurements, is operationalized as a change in perception of the response categories of items in RMT models. In the simulation study, the occurrence of RC was characterized by a difference in threshold parameters between both times.

Table 2
Matrix of threshold parameters at time 1 according to the number of items (J) and number of response categories (M).

	Number of response categories (M)	M = 4	M = 7
Number of items (J)		$\delta_{j1g}^{(1)}$ $\delta_{j2g}^{(1)}$ $\delta_{j3g}^{(1)}$	$\delta_{j1g}^{(1)}$ $\delta_{j2g}^{(1)}$ $\delta_{j3g}^{(1)}$ $\delta_{j4g}^{(1)}$ $\delta_{j5g}^{(1)}$ $\delta_{j6g}^{(1)}$
J = 4	Item 1	−1.84 −0.84 0.16	−1.85 −1.45 −1.05 −0.65 −0.25 0.16
	Item 2	−1.25 −0.25 0.75	−1.26 −0.86 −0.46 −0.06 0.35 0.75
	Item 3	−0.75 0.25 1.25	−0.75 −0.35 0.05 0.46 0.86 1.26
	Item 4	−0.16 0.84 1.84	−0.16 0.25 0.65 1.05 1.45 1.85
J = 7	Item 1	−2.16 −1.16 −0.16	−2.15 −1.75 −1.35 −0.95 −0.55 −0.15
	Item 2	−1.68 −0.68 0.32	−1.68 −1.28 −0.88 −0.48 −0.08 0.32
	Item 3	−1.32 −0.32 0.68	−1.32 −0.92 −0.52 −0.12 0.28 0.68
	Item 4	−1.01 −0.01 0.99	−1.00 −0.60 −0.20 0.20 0.60 1.00
	Item 5	−0.68 0.32 1.32	−0.68 −0.28 0.12 0.52 0.92 1.32
	Item 6	−0.33 0.67 1.67	−0.32 0.08 0.48 0.88 1.28 1.68
	Item 7	0.15 1.15 2.15	0.15 0.55 0.95 1.35 1.75 2.15

$\delta_{jpg}^{(1)}$: p^{th} threshold parameter for group g at time 1 of item j; for each matrix rows are items and column are threshold parameters.

Two types of RC were distinguished: uniform if all the threshold parameters of a given item differed across times, in the same direction and to the same extent, or non-uniform otherwise. The amount of RC of item j for threshold p in group g is defined as the shift in threshold parameters between time 1 and time 2 $\Delta_{jpg} = \delta_{jpg}^{(2)} - \delta_{jpg}^{(1)}$. Fig. 1 illustrates a case of non-uniform RC where the first and the last threshold parameters change across times for a given item. For example, the third threshold parameter increase is $\Delta_{j3g} = 1.5$ between time 1 and time 2 ($\delta_{jpg}^{(1)} = 0.75$ at time 1 and $\delta_{jpg=}^{(1)}$2.25 at time 2). In such a case, an individual with the same level of latent variable (e.g. 2 on the latent variable scale) at both times is more likely to answer $x = 3$ at time 1 and $x = 2$ at time 2 due to RC, operationalized as the increase of the third threshold parameter.

RC could affect none, one or both groups:

- When no RC (noRC) was simulated, all threshold parameters of all items remained the same between the time 1 and time 2. ($\delta_{jpg}^{(2)} = \delta_{jpg}^{(1)}$, $\forall j,p,g$)
- When RC was simulated on a given item j, its threshold parameters were different across times. Simulated RC could differ as follows:
- - Similar RC (RCSim): the same shifts in threshold parameters were simulated in both groups with the same type of RC (i.e. uniform or non-uniform). ($\delta_{jpg}^{(2)} = \delta_{jpg}^{(1)} + \Delta_p \forall p,g$)
- - Differential RC only in one group (RC1grp): RC was simulated only in group 1. ($\delta_{jp1}^{(2)} = \delta_{jp1}^{(1)} + \Delta_p$ *and* $\delta_{jp0}^{(2)} = \delta_{jp0}^{(1)} \forall p$)
- - Differential RC in both groups (RC2grp): RC was simulated in both groups with different shifts for each group but the same type of RC. ($\delta_{jpg}^{(2)} = \delta_{jpg}^{(1)} + \Delta_{jpg} \forall p,g$)

Table 3 presents the sets of shifts in threshold parameters for each group (Δ_{jpg}) according to the scenario of simulated RC (RCSim, RC1grp or RC2grp), the number of response categories and type of RC. The simulated values of threshold parameters at time 2 are equal to $\delta_{jpg}^{(2)} = \delta_{jpg}^{(1)} + \Delta_{jpg}$. For example, when non-uniform similar RC was simulated on item 2 with 4 response categories, the first, second and third threshold parameters were −1.25, −0.25 and 0.75 respectively at time 1 (see Table 2). The components of the set {-1,0,1} were added to the threshold parameters to obtain the simulation values for time 2: −2.25, −0.25, and 1.75 respectively.

Only one item was affected by RC: either a central item (item 2 when J = 4, item 4 when J = 7), or an extreme item, here with the item with the highest threshold parameters (item 4 when J = 4, item 7 when J = 7). The combination of all parameter values led to 624 different cases. 500 replications were simulated for each case.

2.3. ROSALI

ROSALI is based on models from RMT, cross-sectional and longitudinal PCMs.

- Cross-sectional PCM at time 1

The probability for patient i, belonging to group g, to answer x to item j is given by :

$$P\left(X_{ij}^{(1)} = x|\theta_i^{(1)}, \beta, g_i, \delta_{j1g}^{(1)}, \cdots, \delta_{j(M-1)g}^{(1)}\right) = \frac{\exp\left(x\left(\beta \times g_i + \theta_i^{(1)}\right) - \left(\sum_{p=1}^{x} \delta_{jpg}^{(1)}\right)\right)}{\sum_{l=0}^{(M-1)} \exp\left(l\left(\beta \times g_i + \theta_i^{(1)}\right) - \left(\sum_{p=1}^{l} \delta_{jpg}^{(1)}\right)\right)}$$

With $\Theta^{(1)} \sim N\left(\mu^{(1)};\ \sigma_1^2\right)$

- Longitudinal PCM

The probability for patient i, belonging to group g, to answer x to

Table 3
Sets of shifts in threshold parameters for each group according to the scenario of simulated recalibration (RCSim, RC1grp or RC2grp), the number of response categories and the type of recalibration.

		Uniform recalibration		Non-uniform recalibration	
Scenario	Number of response categories	group 0	group 1	group 0	group 1
RCSim	4	{1,1,1}		{-1,0,1}	
	7	{1,1,1,1,1,1}		{-1,-0.5,0,0,0.5,1}	
RC1grp	4	{0,0,0}	{1,1,1}	{0,0,0}	{-1,0,1}
	7	{0,0,0,0,0,0}	{1,1,1,1,1,1}	{0,0,0,0,0,0}	{-1,-0.5,0,0,0.5,1}
RC2grp	4	{-0.8,-0.8,-0.8}	{1,1,1}	{-1,0,0}	{0,0,1.5}
	7	{-0.8,-0.8,-0.8,-0.8,-0.8,-0.8}	{1,1,1,1,1,1}	{-1,-0.5,0,0,0,0}	{0,0,0,0,0.5,1}

RCSim : Similar recalibration, RC1grp : Differential recalibration when RC affects only group 1, RC2grp : Differential recalibration when RC affects both groups.
$\{\Delta_{j1g}, \Delta_{j2g}, \cdots, \Delta_{j(M-1)g}\}$: set of shifts in threshold parameters for each response category (p = 1,…,M − 1) of an item j affected by RC for simulated patients from group g.

Table 4
Identifiability constraints of the models of ROSALI.

	Description	Identifiability constraints	
Part 1	Investigating differences in threshold parameters between groups at time 1		
Model A	Full model: difference in threshold parameters between groups at time 1 on all items	$\mu_0^{(1)} = 0, \sigma_0^{2(1)} = \sigma_1^{2(1)}$	$\mu_0^{(1)} = \mu_1^{(1)}$
Model B	Restricted model: equality of threshold parameters between groups at time 1 on all items	$\mu_0^{(1)} = 0, \sigma_0^{2(1)} = \sigma_1^{2(1)}$	
Model C	Model B with equality constraints relaxed for 1 or more items	$\mu_0^{(1)} = 0, \sigma_0^{2(1)} = \sigma_1^{2(1)}$	
Part 2	Recalibration detection		
Model 1	Full model: recalibration on all items	$\mu_0^{(1)} = 0, \sigma_0^{2(1)} = \sigma_1^{2(1)}, \sigma_0^{2(2)} = \sigma_1^{2(2)}$	$\mu_0^{(1)} = \mu_0^{(2)}, \mu_1^{(1)} = \mu_1^{(2)}, \beta_{inter} = 0$
Model 2	Restricted model: no recalibration	$\mu_0^{(1)} = 0, \sigma_0^{2(1)} = \sigma_1^{2(1)}, \sigma_0^{2(2)} = \sigma_1^{2(2)}$	
Model 3	Model 2 with equality constraints across times relaxed for 1 or more items	$\mu_0^{(1)} = 0, \sigma_0^{2(1)} = \sigma_1^{2(1)}, \sigma_0^{2(2)} = \sigma_1^{2(2)}$	
Model 4		$\mu_0^{(1)} = 0, \sigma_0^{2(1)} = \sigma_1^{2(1)}, \sigma_0^{2(2)} = \sigma_1^{2(2)}$	

$\mu_g^{(t)}$: mean of the latent variable for group g (g = 0,1) at time t (t = 1,2).
$\sigma_g^{2(t)}$: variance of the latent variable for group g (g = 0,1) at time t (t = 1,2).
β_{inter}: coefficient associated with the group by time interaction.

parameters are constrained to be equal between groups. Models A and B are compared with a likelihood ratio test (LRT). If this test is not significant at a significance level of 5%, threshold parameters of all items are constrained to be equal between both groups at time 1 and ROSALI moves on to part 2 for RC detection (see Fig. 3). Otherwise, if the LRT is significant, differences in threshold parameters between both groups are suspected and ROSALI proceeds to step C which is an iterative step to detect which items seem to have different threshold parameters between groups at time 1, starting from model B. At each iteration of step C, constraints of equality of threshold parameters between groups are relaxed item-by-item. For each item, the equality of threshold parameters for all response categories between groups is tested with a Wald test. A Bonferroni [25] correction accounting for the number of items to be tested is applied to avoid inflation of the type I error rate due to multiple testing. Item with the most significant test is selected and the equality of the difference in threshold parameters across all response categories is tested to determine if the difference is uniform or non-uniform for this item. The model is updated to take account of (non–)uniform differences in threshold parameters between groups on the selected item *j*, and step C is repeated on this updated model to identify differences on the remaining items.

This step is stopped when there are no items left with a significant difference between groups at time 1 or when only one item remains to be tested and ROSALI moves on to part 2.

2.3.2. Part 2 of ROSALI: RC detection

Fig. 3 presents the different steps of the second part of ROSALI. The second part focuses on the detection of difference in threshold parameters across times. Different longitudinal PCMs are used in this part, taking into account differences in threshold parameters between groups

$$P\left(X_{ij}^{(t)} = x|\theta_i^{(t)}, \beta, g_i, \beta_{inter}, t_2, \delta_{j1g}^{(t)}, \cdots, \delta_{j(\mathrm{M}-1)g}^{(t)}\right) = \frac{\exp\left(x\left(\beta \times g_i + \beta_{inter} \times t_2 \times g_i + \theta_i^{(t)}\right) - \left(\sum_{p=1}^{x} \delta_{jpg}^{(t)}\right)\right)}{\sum_{l=0}^{(\mathrm{M}-1)} \exp\left(l\left(\beta \times g_i + \beta_{inter} \times t_2 \times g_i + \theta_i^{(t)}\right) - \left(\sum_{p=1}^{l} \delta_{jpg}^{(t)}\right)\right)}$$

item *j* at time *t* is given by :

$$\text{With } \begin{bmatrix} \Theta^{(1)} \\ \Theta^{(2)} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}; \Sigma\right)$$

The model parameters are estimated using marginal maximum likelihood. One identifiability constraint is applied to all models of ROSALI: the nullity of the mean of the latent variable at time 1 for group 0 ($\mu_0^{(1)} = 0$). Other identifiability constraints are applied according to the steps of ROSALI detailed below and are summarized in Table 4.

ROSALI consists of 2 main parts [17]. In the first part (steps A to C), differences in threshold parameters between groups at time 1 are investigated. The second part (steps 1 to 4) consists in the detection of items affected by RC and the estimation of the group effects on the latent variable and of RC effects. ROSALI is presented in Figs. 2 and 3. The different steps of ROSALI are summarized below. Full details on the ROSALI algorithm can be found elsewhere [17]. ROSALI has been automated into a Stata (*Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC) module stored at Boston College's Statistical Software Components archive [24].

2.3.1. Part 1 of ROSALI: Investigating differences in threshold parameters between groups at time 1

Fig. 2 presents the different steps of the first part of ROSALI. Different cross-sectional PCMs at time t = 1 are used in the first part: a model A (full model) where all threshold parameters of all items are freely estimated between groups and a restricted model B where all threshold

at time 1 identified in the first part (steps A to C). The first step is to compare two models with a LRT: a full model 1 where all threshold parameters of all items are freely estimated across times (RC on all items) and a restricted model 2 where all threshold parameters are constrained to be equal across times (no RC on any item). If the LRT is not significant at 5%, all threshold parameters are constrained to be equal across times and ROSALI moves on to step 4. Otherwise, if the LRT is significant, differences in threshold parameters across times (RC) are suspected and ROSALI proceeds to the iterative step 3, to identify which items seem to be affected by RC. For each item, starting from model 2, several models 3 are estimated where constraints of equality of threshold parameters across times are relaxed item-by-item. For each item j, the equality of threshold parameters for all response categories between time 1 and time 2 is tested with a Wald test of overall RC occurrence. A Bonferroni correction accounting for the number of items to be tested is applied. The item with the most significant test is selected. A Wald test is realized to see if RC on the selected item is similar in both groups or not :

- If the Wald test is significant: differential RC is assumed and occurrence of RC is tested group-by-group with a Wald test and a Bonferroni adjustment according to the number of groups is applied. Then, for each group where RC is detected, a test evaluates whether RC is considered uniform or not.
- If the Wald test is not significant: similar RC is assumed, model 3 is updated to add a constraint of similar RC for the selected item. Then,

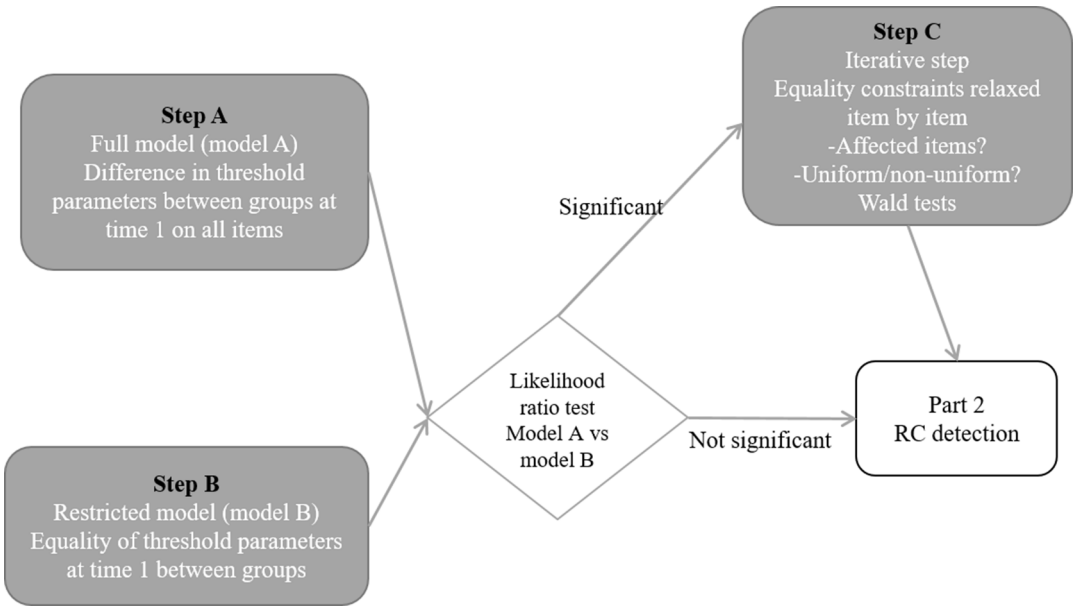

Fig. 2. Part 1 of ROSALI (Step A–C). Cross-sectional Partial Credit Models at time 1 to detect differences in threshold parameters between groups. RC: recalibration.

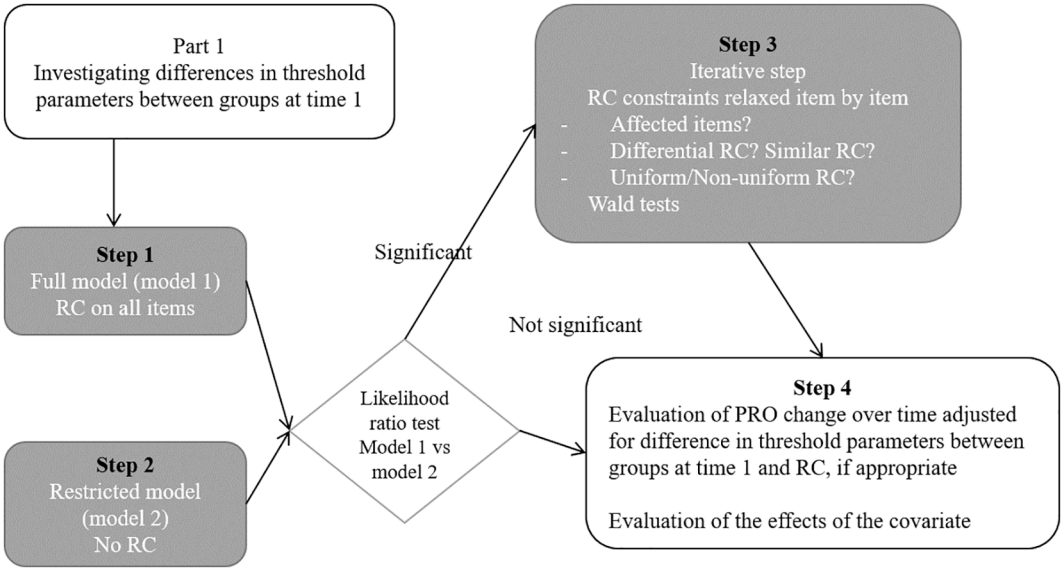

Fig. 3. Part 2 of ROSALI (Step 1–4). Longitudinal Partial Credit Models to detect difference in threshold parameters across times (recalibration, RC) and evaluation of covariate's effects.

a Wald test is performed on the estimations of the updated model to evaluate the type of RC on this item (uniform or non-uniform).

Model 3 is updated to account for differences in threshold parameters across times on the selected item, and step 3 is repeated on this updated model to identify RC on the remaining items. This step is stopped when there are no items left with a significant RC test or when only one item remains to be tested.

A final step 4 is performed. A model 4 is estimated accounting for differences in threshold parameters between groups at time 1 found in part 1 and differences in threshold parameters across times (RC) found in part 2, if appropriate. Effects on the latent variable means are estimated and tested with Wald tests: group effect (H_0: $\beta = 0$), time effect ($H_0 : \mu_0^{(2)} = 0$) and time × group interaction ($H_0 : \beta_{inter} = 0$). Differences in threshold parameters between groups or/and RC effects are also estimated in model 4.

3. Analysis

ROSALI was applied to each replicated dataset of the 624 simulated cases. ROSALI performance was assessed using model 4 in terms of RC detection and bias in the estimations of the parameters related to the latent variable. Datasets were simulated and analyzed with Stata (*Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC).

3.1. Evaluation of the performance of ROSALI

For datasets in which no RC was simulated (noRC), the rate of false detection of RC was defined as the percentage of datasets for which RC was detected on at least one item. Likewise, the rate of false detection of differences in threshold parameters between groups at time 1 was also computed, on these datasets, since no group difference was simulated at time 1.

When RC was simulated (RCSim, RC1grp, and RC2grp), different

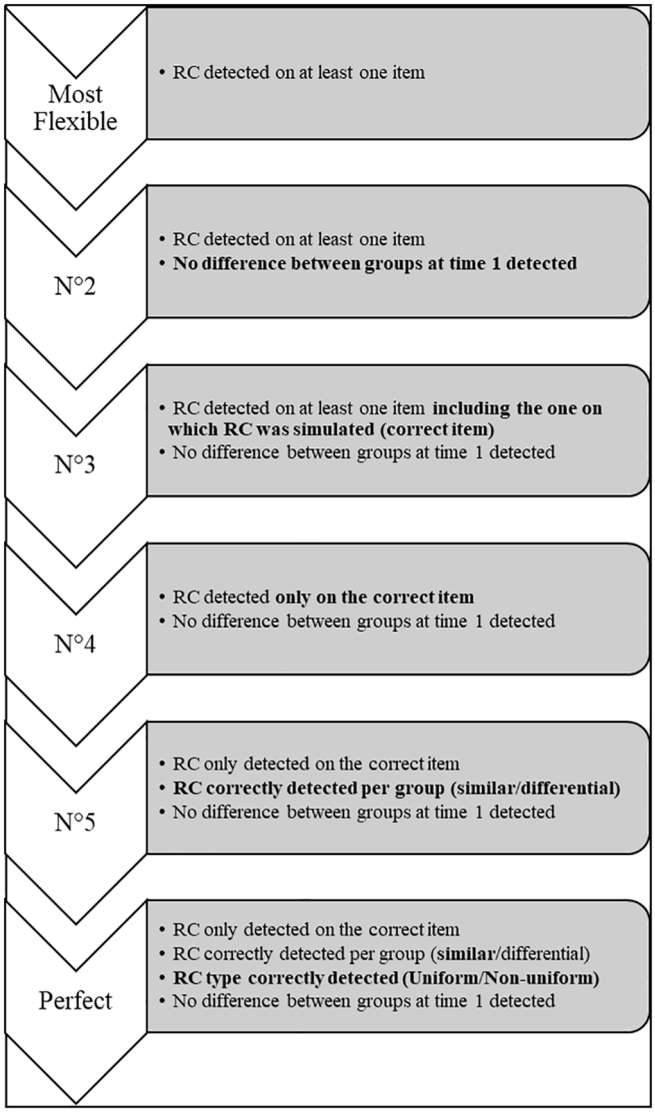

Fig. 4. Criteria for evaluation of ROSALI algorithm. The condition in bold indicates the additional condition when moving from one criteria to the other from "Most flexible" to "Perfect". RC: recalibration.

criteria were defined to assess the performance of RC detection and are presented in Fig. 4. Each criterion, increasingly restrictive, was established to investigate the performance of the different steps of ROSALI.

- Criterion n°1 – most flexible: Was ROSALI able to detect RC on at least one item (including or not the item on which RC was simulated)?
- Criterion n°2: Was ROSALI able to identify the absence of difference in threshold parameters between groups at time 1?
- Criterion n°3: Was ROSALI able to rightly detect RC on at least the item on which RC was simulated?
- Criterion n°4: Was ROSALI able to rightly detect RC only on the item on which RC was simulated?
- Criterion n°5: Was ROSALI able to rightly identify similar or differential RC and the groups in which RC was simulated?
- Criterion n°6 - perfect: Was ROSALI able to rightly detect the type of RC (uniform or non-uniform)? That is, was ROSALI able to detect exactly what was simulated (item on which RC was simulated, similar or differential RC, group(s) on which RC was simulated and type of RC, and nothing else)?

The performance of ROSALI was assessed using the percentage of datasets that met the different criteria (from Most flexible to Perfect). The difference between the percentages of datasets meeting the different criteria was also studied. For instance, the difference between criteria n°2 and n°3 indicates how much ROSALI failed to identify the correct item on which RC was simulated. These criteria aimed to help the diagnosis of the strengths and weaknesses of the algorithm. Thus, ROSALI can be considered as having good performance if the percentages of datasets meeting the different criteria are high and differences between criteria are low.

3.2. Type I error and power of the tests of group, time, and interaction effects

When group and time effect as well as their interaction were simulated at 0, the type I error of the test of each effect was calculated as the percentage of datasets where the test of the nullity of the associated parameter was significant. Besides, when group and time effects were simulated as being different from 0, the power of the tests was computed as the percentage of datasets where the tests were significant. Biases in

Table 5
Detection rates using the most flexible and perfect criteria, according to: scenarios of simulated recalibration (RCSim, RC1grp, RC2grp), number of items (J), number of response categories (M) and sample size (N) for datasets with simulated (non–)uniform recalibration.

				Uniform recalibration		Non-uniform recalibration	
Scenario	J	M	N	Most flexible (%)	Perfect (%)	Most flexible (%)	Perfect (%)
RCSim	4	4	200	98	84	61	53
			300	100	86	86	76
			500	100	85	99	88
	4	7	200	100	89	53	47
			300	100	88	83	75
			500	100	87	99	89
	7	4	200	95	83	50	42
			300	100	85	78	69
			500	100	86	98	88
	7	7	200	100	89	45	39
			300	100	88	75	68
			500	100	87	97	89
RC1grp	4	4	200	92	80	42	32
			300	99	87	71	60
			500	100	87	96	87
	4	7	200	100	90	33	25
			300	100	90	62	53
			500	100	89	93	85
	7	4	200	87	76	32	25
			300	98	85	60	50
			500	100	89	90	82
	7	7	200	98	90	25	19
			300	100	90	51	45
			500	100	89	87	79
RC2grp	4	4	200	100	77	62	11
			300	100	85	87	31
			500	100	84	100	61
	4	7	200	100	89	31	3
			300	100	88	62	10
			500	100	86	94	34
	7	4	200	100	82	49	11
			300	100	87	77	29
			500	100	86	98	63
	7	7	200	100	90	47	7
			300	100	88	80	24
			500	100	87	97	52

The results are summarized for all simulated values of group effect, time effect and item position.
RCSim: Similar recalibration, RC1grp: Differential recalibration with recalibration only in one group, RC2grp: Differential recalibration with recalibration in both groups, J: Number of items in the d, M: Number of response categories, N: Sample size, NU: Non-uniform, U: Uniform, RC: recalibration.

the estimation of group effect, time effect and interaction between group and time were estimated as the difference between the mean of the estimations of the effect and the value of the corresponding simulation parameter in each of the 624 different cases.

4. Results

4.1. No RC simulated

Rates of false detection of a difference in threshold parameters between groups at time 1 and/or of RC in model 4 were low, ranging from 1.6% to 7.0% (results not shown). None of the simulation parameters (sample size, time effect, group effect, number of items and response categories) seemed to have an impact on this rate.

4.2. RC simulated

Table 5 presents detection rates using the most flexible and perfect criteria according to different scenarios of simulated RC (RCSim, RC1grp or RC2grp), number of items (J), number of response categories (M) and sample size (N). Table 6 presents differences in rates between the increasingly restrictive criteria. Group effect, time effect, and the item position (central or extreme) did not seem to have an influence on these rates. Hence, the following results are summarized for all simulated values of group effect, time effect and item position.

4.2.1. Performance of ROSALI when uniform RC was simulated

Rates of detection using the most flexible criterion in Table 5 were high (they ranged between 87% and 100%), meaning that ROSALI was able to detect RC on at least one item but which may not include the item on which uniform RC was simulated. Detection rates using the perfect criterion were also quite high (they ranged between 76% and 90%) meaning that ROSALI usually correctly detected what was exactly simulated (item with RC, type of RC, scenarios of simulated RC: RCSim, RC1grp or RC2grp).

The difference in rates between the increasingly restrictive criteria, presented in Table 6, can give us some additional clues on the performance of ROSALI. For instance, the differences between the rates related to criteria n°4 and n°5 were usually low and showed that the scenarios of simulated RC (RCSim, RC1grp or RC2grp) were correctly identified in most cases (differences ranging from 0% to 5%), except when differential RC was simulated in 2 groups (RC2grp) and J = 4, M = 4 and N =

Table 6
Differences in detection rates between between criteria, according to the groups affected by recalibration, number of items (J), number of response categories (M) and sample size (N) for datasets with simulated (non-)uniform recalibration.

Difference between criteria				100% – Most flexible (%) RC not detected		Most flexible – n°2 (%) Differences between groups at time 1 detected		n°2 – n°3 (%) Item with simulated RC not detected		n°3 – n°4 (%) Item with simulated RC detected + other items		n°4 – n°5 (%) Wrong group detected/wrong scenario		n°5 – Perfect (%) Wrong type of RC	
Scenario	J	M	N	U	NU	U	NU	U	NU	U	NU	U	NU	U	NU
RCSim	4	4	200	2	39	3	2	0	1	3	2	4	3	4	0
			300	0	14	2	2	0	0	4	3	4	4	4	0
			500	0	1	2	2	0	0	4	4	4	4	4	0
	4	7	200	0	47	2	1	0	1	3	2	3	3	4	0
			300	0	17	2	2	0	0	3	3	3	4	4	0
			500	0	1	2	2	0	0	3	4	3	4	4	0
	7	4	200	5	51	2	1	0	0	3	3	4	3	4	1
			300	0	22	3	1	0	0	4	4	4	4	4	0
			500	0	2	2	2	0	0	4	4	4	4	4	0
	7	7	200	0	55	1	0	0	0	2	2	3	2	5	1
			300	0	25	1	1	0	0	3	3	3	3	4	0
			500	0	3	2	1	0	0	4	4	3	3	4	0
RC1grp	4	4	200	8	58	2	1	0	1	3	2	3	4	4	1
			300	1	29	2	2	0	1	4	3	2	5	4	0
			500	0	4	3	2	0	0	4	4	2	3	4	0
	4	7	200	0	67	2	1	0	1	3	1	2	4	3	1
			300	0	38	2	1	0	1	3	2	2	4	4	1
			500	0	7	2	2	0	0	4	4	2	3	4	0
	7	4	200	14	68	2	1	0	1	3	1	3	3	3	1
			300	2	40	2	1	0	1	4	3	2	4	4	0
			500	0	10	2	2	0	0	4	4	2	3	3	0
	7	7	200	2	75	1	0	0	1	2	1	1	2	3	1
			300	0	49	2	1	0	1	3	2	2	2	3	1
			500	0	13	2	1	0	0	4	3	2	2	4	0
RC2grp	4	4	200	0	38	2	1	0	1	4	3	9	36	7	9
			300	0	13	2	1	0	0	4	4	1	38	8	13
			500	0	0	2	2	0	0	4	4	0	21	9	11
	4	7	200	0	69	1	1	0	2	3	1	0	18	7	7
			300	0	38	2	1	0	1	4	3	0	31	7	16
			500	0	6	2	2	0	0	3	4	0	26	8	27
	7	4	200	0	51	1	1	0	1	4	2	5	24	7	10
			300	0	23	2	1	0	0	4	3	0	28	8	15
			500	0	2	2	1	0	0	4	4	0	14	8	16
	7	7	200	0	53	1	1	0	0	2	2	0	28	7	10
			300	0	20	1	1	0	0	3	3	0	33	7	19
			500	0	3	1	1	0	0	4	4	0	16	9	24

The results are summarized for all simulated values of group effect, time effect and item position.
Simulated scenarios: RCSim: Similar recalibration, RC1grp: Differential recalibration with recalibration only in one group, RC2grp: Differential recalibration with recalibration in both groups.
J: Number of items in the domain, M: Number of response categories, N: Sample size, NU: Non-uniform, U: Uniform, RC: recalibration.

200 (differences between criteria n°4 and n°5 was 9% on average). The correct type of RC was identified in case of similar RC or when RC was simulated in one group only (differences between the detection rates using criterion n°5 and the perfect criterion, ranged from 3% to 5% for RCSim and RC1grp). However, differences between these criteria were a bit higher (ranging from 7% to 9%) when differential RC was simulated in both groups (RC2grp) meaning that ROSALI assumed that RC was non-uniform instead of uniform.

4.2.2. Performance of ROSALI when non-uniform RC was simulated

In Table 5, the rates regarding the most flexible criterion ranged from 25% to 100%. These rates increased with sample size and decreased as the number of items and of response categories increased especially when the sample size was lower than 300. It indicates that ROSALI struggled in identifying RC when the sample size was equal to 200 and it was even harder as the number of items or response categories increased. Overall, lower detection rates using the most flexible criterion were more often observed when differential RC was simulated in one group only (RC1grp) meaning that ROSALI had more difficulty detecting RC in this scenario compared to scenarios of similar RC (RCSim) or RC simulated in both groups (RC2grp). Detection rates using the perfect criterion also increased with sample size and decreased as the number of items and of response categories increased. These rates were the highest when similar RC was simulated (RCSim, range: 32%-89%) and the lowest when differential RC was simulated in both groups (RC2grp scenario, range: 3–63%).

The difference in rates between the increasingly restrictive criteria for simulated non-uniform RC, in Table 6, show that generally, the item and only the items on which RC was simulated was correctly detected (difference between criteria n°2 and n°4 ranged from 2% to 4%) provided that ROSALI detected RC on at least one item. When similar RC (RCSim) or differential RC in only one group (RC1grp) was simulated, the group(s) was (were) often well-identified (difference between criteria n°4 and n°5 ranged from 2% to 5%), as well as the correct type of simulated RC (difference between criteria n°5 and n°6 ranged from 0% to 1%). However, when differential RC in both groups (RC2grp) was simulated, ROSALI often identified similar RC or RC in one group only instead of differential RC in both groups (difference between criteria n°4 and n°5 ranged from 14% to 38%). ROSALI also more frequently detected the wrong type of RC (difference between criteria n°5 and n°6 ranged from 7% to 27%) meaning that ROSALI assumed that RC was uniform instead of non-uniform.

4.2.3. Estimations of group effect, time effect and time × group interaction

Table 7 presents the type I error of the Wald tests of group effect, time effect, and their interaction when they were simulated at 0 and power of the Wald tests of group and time effect when they were simulated as different from 0. Results are presented for the final model of ROSALI accounting for differences in threshold parameters between groups and over time, i.e. model 4, and for the model assuming no RC, i.e. model 2.

In model 4, the type I error was usually well controlled when no group, time effects or their interaction were simulated, (group effect: 5% to 7%, time effect: 6% to 9% and interaction time × group: 5% to 8%). When a group effect was simulated, the power for the test of group effect ranged between 29% and 64%. When a time effect was simulated, the estimated power for the test of time effect was ranged between 91% and 100%. The power of each test increased with sample size, as expected.

4.2.4. Adjustment of group effect, time effect and time × group interaction for RC

The occurrence of RC can bias the estimated means of the latent variable. To give some indications on the ability of ROSALI to correctly adjust estimated means for RS, estimated bias of group and time effects (Fig. 5) as well as type I error and power of the tests of group and time effects, and interaction can be compared between model 2 not accounting for RC and model 4 accounting for RC. On Fig. 5, estimations of group effect were unbiased in model 2 and in model 4. Estimations of time effect were biased only in model 2 when uniform RC was similarly (RCSim) or differentially simulated in both groups (RC2grp) or when non-uniform RC was simulated differentially in both groups (RC2grp). Estimations of time × group interaction were biased only in model 2 when uniform RC was differentially simulated in one or both groups (RC1grp or RC2grp) or when non-uniform RC was simulated differentially in both groups (RC2grp).

Only small differences were observed on type I error of the test of group effect between model 2 and model 4 (Table 7). However, the type I error of the test of time effect, in model 2, were very high (range: 28%-65%) and much higher than in model 4 when uniform RC was simulated similarly (RCSim) or differentially in both groups (RC2grp). Similarly, the type I error of the interaction test was also much higher in model 2 (range: 23%-86%) as compared to model 4 (range: 5%-6%) when uniform differential RC (RC1grp or RC2grp) was simulated.

For the tests of group effect, the estimations of statistical power were always lower in model 2 than in model 4. The difference between the estimated power ranged from 4% to 12%. For the tests of time effect, the estimations of statistical power were usually lower in model 2 than in model 4. The difference between the estimated power ranged from − 3% to 68%. It was the highest when uniform RC was simulated similarly in the two groups (RCSim) and it decreased as sample size increased, for all cases.

5. Discussion

This simulation study assessed the performances of ROSALI in terms of RC detection and bias in the estimations of the parameters related to the latent variable in the context of longitudinal studies designed for the comparison of two groups of patients. Rates of false detection of RC and/or difference in threshold parameters between groups at time 1 were low indicating that ROSALI satisfactorily prevents from erroneously inferring a difference in threshold parameters between groups or across times. These good performances may be related to the Bonferroni correction applied in the two iterative steps (step C and step 3) following likelihood ratio tests and to the iterative steps themselves. This asset of ROSALI has already been shown in a previous simulation study [14]. When the LRT erroneously suggests the presence of a RC or a difference in threshold parameters between groups at time 1, this error is often corrected by iterative steps 3 and C, respectively, within which no items are flagged.

In the presence of uniform RC, ROSALI is able to detect RC, identify the item and the group(s) affected by RC and the type of RC in light of the high levels of detection rates using the perfect criterion. However, these detection rates were a bit lower when uniform RC was simulated differentially in two groups as ROSALI struggled with the identification of the groups affected by RC. For almost all cases, the size of shift in threshold parameters for uniform RC was + 1 but for differential RC in two groups, the size was − 0.8 in one group and + 1 in the other group, making RC likely harder to detect. It therefore seems that we can be confident that a uniform RC greater than 1 can be detected by ROSALI for studies with similar RC and sample sizes in each group, number of items and response categories.

Performances of ROSALI for RC detection was usually lower for non-uniform RC than uniform RC. Indeed, non-uniform RC was sometimes difficult to detect at onset (low detection rates using the most flexible criterion, especially for sample sizes lower than 300). However, as soon as ROSALI could identify non-uniform RC on at least one item, the item and only the item on which it had been simulated was correctly detected. When similar RC or differential RC in only one group was simulated, the group(s) was (were) often well-identified, as well as the correct type of simulated RC, contrary to differential RC in both groups. The poor performances for non-uniform RC may be linked to the values of the simulated shifts in threshold parameters. For example, the threshold

Table 7
Type I error of the tests of group effect, time effect and their interaction in models 2 and 4, and power of the tests of group and time effect on models 2 and 4 according to type of RC, simulated scenario (RCSim, RC1grp, RC2grp) and sample size.

			Type I error						Power			
			Group effect = 0		Time effect = 0		Interaction time × group = 0		Group effect = 0.2		Time effect = 0.3	
		N	Not adjusted for RC	Adjusted for RC	Not adjusted for RC	Adjusted for RC	Not adjusted for RC	Adjusted for RC	Not adjusted for RC	Adjusted for RC	Not adjusted for RC	Adjusted for RC
No RC		200	5%	6%	5%	7%	5%	5%	25%	33%	75%	94%
		300	5%	6%	5%	6%	5%	5%	35%	44%	89%	98%
		500	5%	6%	5%	6%	5%	5%	52%	64%	98%	99%
Uniform	RC Sim	200	5%	6%	35%	8%	5%	5%	25%	32%	23%	91%
		300	5%	6%	47%	7%	6%	6%	36%	44%	29%	97%
		500	5%	6%	65%	7%	5%	5%	52%	64%	41%	99%
	RC1grp	200	6%	7%	5%	7%	23%	6%	26%	33%	75%	92%
		300	5%	6%	5%	7%	31%	5%	34%	44%	90%	97%
		500	4%	5%	5%	6%	47%	5%	52%	63%	98%	100%
	RC2grp	200	5%	6%	28%	7%	56%	5%	25%	32%	98%	91%
		300	5%	6%	40%	7%	71%	5%	34%	42%	100%	97%
		500	5%	5%	59%	7%	86%	5%	51%	63%	100%	99%
Non-Uniform	RCSim	200	5%	6%	6%	8%	5%	5%	25%	32%	76%	92%
		300	5%	6%	7%	7%	5%	5%	35%	44%	89%	97%
		500	5%	5%	7%	7%	5%	5%	52%	63%	98%	99%
	RC1grp	200	5%	6%	5%	7%	5%	5%	25%	33%	76%	93%
		300	5%	6%	5%	7%	6%	6%	34%	44%	89%	97%
		500	5%	6%	5%	7%	6%	5%	52%	63%	98%	99%
	RC2grp	200	5%	7%	8%	9%	10%	8%	25%	29%	86%	94%
		300	5%	7%	9%	8%	12%	7%	35%	41%	96%	98%
		500	5%	6%	11%	7%	17%	6%	51%	62%	100%	99%

RCSim: Similar recalibration, RC1grp: Differential recalibration with recalibration only in one group, RC2grp: Differential recalibration with recalibration in both groups, N: sample size.
Not adjusted for RC: model 2 in ROSALI with no RC assumed, Adjusted for RC: final model of ROSALI in which differences in threshold parameters at time 1 and RC are accounted for.

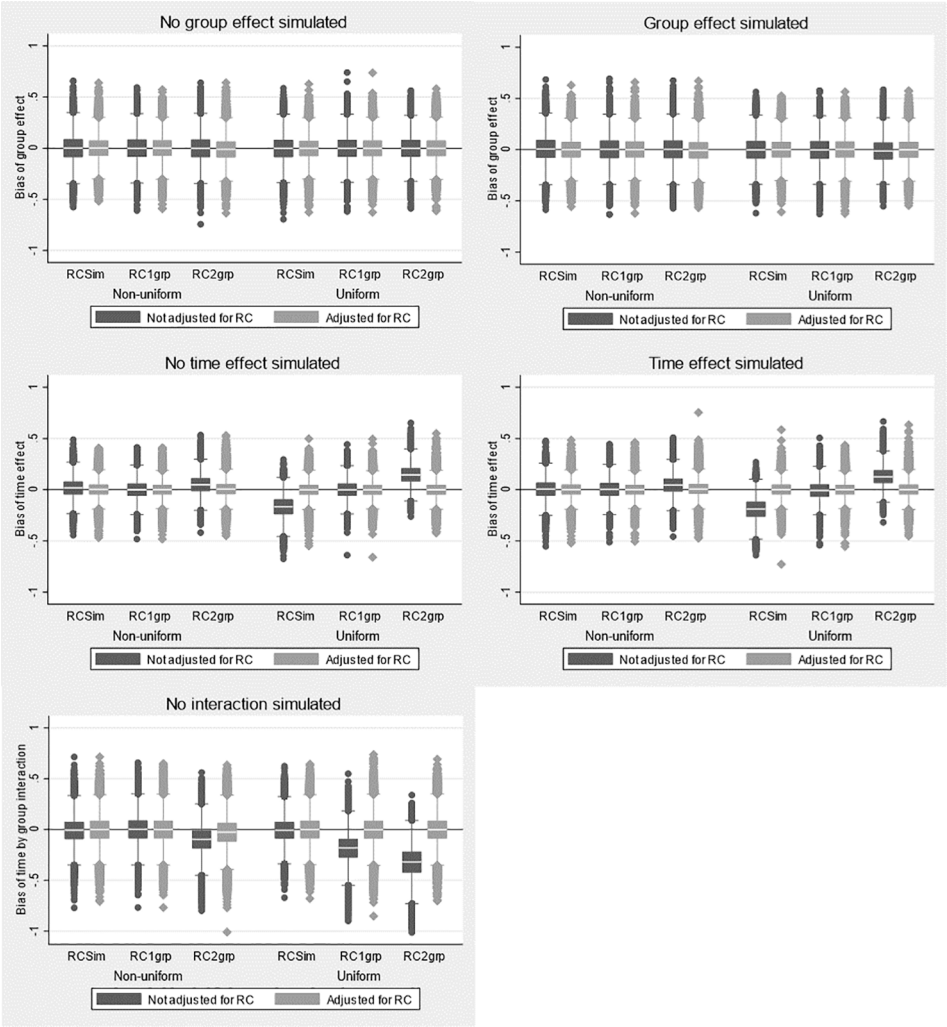

Fig. 5. Bias of group and time effects and time × group interaction when effects were not simulated and when they were simulated according to type of recalibration (RC) and groups affected by RC (RCSim : Similar recalibration, RC1grp : Differential recalibration when RC affects only group 1, RC2grp : Differential recalibration when RC affects both groups). The boxplots compare estimations from model 2 (Not adjusted for RC) and model 4 (Adjusted for RC) from ROSALI.

parameters shifted of − 1, 0, and 1 in case of non-uniform RC for an item with 4 response categories in the scenarios corresponding to similar RC or differential RC in one group. Only some threshold parameters are affected by RC and the item location stayed the same. On the contrary, the threshold parameters shifted of 1, 1, and 1 in case of uniform RC also shifting the item location to the right. Hence, RC affected less response categories and to a lesser extent in case of non-uniform RC making it more difficult to detect than uniform RC.

The detection rates using the perfect criterion increased with the sample size (N). The effect of sample size seems higher in case of non-uniform RC. However, the detection rates showed a ceiling effect and we can expect that for smaller sizes of uniform RC than in this simulation study, the effect of sample size would be as high as for non-uniform RC. Hence, a sample size higher than 200 is recommended as detection rates using the perfect criteria were not satisfactory for non-uniform RC in datasets with N = 200.

The type I error of the test of group effect was usually well controlled but the power of the test of group effect was low confirming that an effect size of 0.2 for the group effect is small and consequently rather difficult to detect, a fortiori when the sample size is small (N = 200).

Not accounting for RC had an impact on type I error and power of the test of time effect as the type I error of time effect was not well controlled. These results may be explained from the under-estimation of the latent variable mean when similar uniform RC was simulated or its over-estimation when uniform or non-uniform RC was simulated differentially in both groups.

5.1. Limitations and perspectives

The performances of ROSALI were evaluated only in terms of RC detection in this simulation study. The properties of ROSALI in terms of detection of differences in threshold parameters at time 1 could not be studied as no differences between groups have been simulated at time 1. Furthermore, the biases of estimated time by group interaction, and power of the test of interaction could not be estimated as a single simulated value of interaction (=0) was investigated. As for time and group effects, we can expect that the bias of interaction would depend on the size of the interaction term, on the type of RC (uniform or not) and whether the groups are affected the same way by RC. Unbiased estimations of group and time effects were obtained in model 4 adjusting for RC and the same could be expected for the interaction term. The effect of unequal group sizes has not been investigated either. We can expect that the group size in combination with the size of RC will influence the performances of ROSALI when RC occurs only in one group (RC1grp) or differentially in both groups (RC2grp). Other conditions have to be studied in the near future to complete the assessment of the performances of ROSALI. All simulated datasets were complete meaning that there were no missing data with regards to item responses nor group

covariate. A simulation study could inform on the impact of missing data on the performances of ROSALI. As parameters are estimated using marginal maximum likelihood, it is expected that in case of MCAR and MAR mechanisms and intermittent missing items or missing group covariates, the estimations will be asymptotically unbiased. However, given the loss of precision of estimators, a loss in the performances of ROSALI can be expected. Furthermore, a pre-requisite to the use of ROSALI is that a unidimensional PCM should fit the data at each time. Robustness of ROSALI in case of deviations from the underlying assumptions of the PCM is unknown and could be investigated in further simulation studies. For instance, a model from item response theory like a longitudinal generalized partial-credit model (GPCM) might fit the data better. Depending on the amount of misfit, the estimators might be biased [26] resulting in a loss of performances for ROSALI. Model misfit may also be indicative of RS. If a GPCM better fits the data at time 2 only, it can be indicative of reprioritization RS operationalized as a change in discrimination parameters (constrained to be equal to 1 in the PCM but freely estimated in the GPCM). If a multidimensional model better fits the data at time 2, the definition of the target construct may have changed indicating reconceptualization RS. Last, models from Rasch measurement theory (e.g. the PCM) assumes local independence, i.e. the items are conditionally independent given the latent variable. In the longitudinal PCM of ROSALI, the items are also assumed to be locally independent across time points as the latent variables are correlated at each time point but item responses are not correlated. The violation of this assumption can also cause misfit of the PCM. Local dependence of an item over time is operationalized as a change in threshold parameters at time 2 depending on the answer at time 1. If a majority of people have answered the same response category at a locally dependent item over time at time 1, RS can be mistakenly suspected in place of local dependence [27]. Local dependence can be one of the alternative explanations of change of threshold parameters for RS detection methods based on Rasch models. Whatever the RS detection methods, possible alternative explanations should be examined to interpret the results [12].

Further developments are needed to better grasp the determinants of RS. Indeed, to take into account the clinical and psychological characteristics of patients, RS detection methods should be able to investigate the effects of covariates with more than two response categories and to investigate the effects of several covariates simultaneously. Finally, ROSALI is relevant when a major health event (e.g. diagnosis, treatment initiation, major surgical operation) has been identified and the analysis is performed before and after this event. However, in chronic conditions, the event that may trigger RS may not occur at the same time for all patients or the time of the event may not be identified prior to the RS analysis. Thus, ROSALI and methods for RS detection in general could benefit from performing the RS detection over multiple time points to help better understanding the psychological adaptation process to chronic conditions. A combination of linear mixed model with RMT by considering the latent variable as a latent process in continuous time may help investigating longitudinal measurement invariance [28]. Indeed, the trajectory of the latent variable over time could be modelled jointly with the trajectories of threshold parameters that may be related to RS.

CRediT authorship contribution statement

Myriam Blanchin: Conceptualization, Software, Writing – original draft. **Priscilla Brisson:** Software, Investigation, Writing – original draft. **Véronique Sébille:** Conceptualization, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study has received funding from the European Union's Horizon 2020 Research and Innovation Program under grant agreement No 847782. The analysis of simulated datasets with ROSALI was performed thanks to the Centre de Calcul Intensif des Pays de la Loire (CCIPL) resources.

References

[1] V.K. Noonan, A. Lyddiatt, P. Ware, S.B. Jaglal, R.J. Riopelle, C.O. Bingham, S. Figueiredo, R. Sawatzky, M. Santana, S.J. Bartlett, S. Ahmed, Montreal accord on patient-reported outcomes (PROs) use series – Paper 3: patient-reported outcomes can facilitate shared decision-making and guide self-management, J. Clin. Epidemiol. 89 (2017) 125–135, https://doi.org/10.1016/j.jclinepi.2017.04.017.

[2] C.O. Bingham, V.K. Noonan, C. Auger, D.E. Feldman, S. Ahmed, S.J. Bartlett, Montreal accord on patient-reported outcomes (PROs) use series – Paper 4: patient-reported outcomes can inform clinical decision making in chronic care, J. Clin. Epidemiol. 89 (2017) 136–141, https://doi.org/10.1016/j.jclinepi.2017.04.014.

[3] C.M. Kieffer, A.R. Miller, B. Chacko, A.S. Robertson, FDA reported use of patient experience data in 2018 drug approvals, Ther. Innov. Regul. Sci. 54 (2020) 709–716, https://doi.org/10.1007/s43441-019-00106-1.

[4] C.-J. Hsiao, C. Dymek, B. Kim, B. Russell, Advancing the use of patient-reported outcomes in practice: understanding challenges, opportunities, and the potential of health information technology, Qual. Life Res. 28 (6) (2019) 1575–1583, https://doi.org/10.1007/s11136-019-02112-0.

[5] A. Vanier, F.J. Oort, L. McClimans, N. Ow, B.G. Gulek, J.R. Böhnke, M. Sprangers, V. Sébille, N. Mayo, Response Shift - in sync working group, response shift in patient-reported outcomes: definition, theory, and a revised model, Qual. Life Res. 30 (12) (2021) 3309–3322, https://doi.org/10.1007/s11136-021-02846-w.

[6] M.A.G. Sprangers, C.E. Schwartz, Integrating response shift into health-related quality of life research: a theoretical model, Soc. Sci. Med. 48 (11) (1999) 1507–1515, https://doi.org/10.1016/S0277-9536(99)00045-3.

[7] M.G.E. Verdam, F.J. Oort, M.A.G. Sprangers, Using structural equation modeling to investigate change and response shift in patient-reported outcomes: practical considerations and recommendations, Qual. Life Res. 30 (2021) 1293–1304, https://doi.org/10.1007/s11136-020-02742-9.

[8] I.D. Hartog, D.L. Willems, W.B. van den Hout, M. Scherer-Rath, T.H. Oreel, J.P. S. Henriques, P.T. Nieuwkerk, H.W.M. van Laarhoven, M.A.G. Sprangers, Influence of response shift and disposition on patient-reported outcomes may lead to suboptimal medical decisions: a medical ethics perspective, BMC Med Ethics 20 (2019) 61, https://doi.org/10.1186/s12910-019-0397-3.

[9] L.M. Lix, E.K.H. Chan, R. Sawatzky, T.T. Sajobi, J. Liu, W. Hopman, N. Mayo, Response shift and disease activity in inflammatory bowel disease, Qual. Life Res. 25 (7) (2016) 1751–1760, https://doi.org/10.1007/s11136-015-1188-z.

[10] M. Salmon, M. Blanchin, C. Rotonda, F. Guillemin, V. Sébille, Identifying patterns of adaptation in breast cancer patients with cancer-related fatigue using response shift analyses at subgroup level, Cancer Med. 6 (11) (2017) 2562–2575, https://doi.org/10.1002/cam4.1219.

[11] T.T. Sajobi, R. Brahmbatt, L.M. Lix, B.D. Zumbo, R. Sawatzky, Scoping review of response shift methods: current reporting practices and recommendations, Qual. Life Res. 27 (5) (2018) 1133–1146, https://doi.org/10.1007/s11136-017-1751-x.

[12] V. Sébille, L.M. Lix, O.F. Ayilara, T.T. Sajobi, A.C.J.W. Janssens, R. Sawatzky, M.A. G. Sprangers, M.G.E. Verdam, Response Shift – in Sync Working Group, Critical examination of current response shift methods and proposal for advancing new methods, Qual. Life Res. 30 (12) (2021) 3325–3342, https://doi.org/10.1007/s11136-020-02755-4.

[13] C.E. Schwartz, Introduction to special section on response shift at the item level, Qual. Life Res. 25 (6) (2016) 1323–1325, https://doi.org/10.1007/s11136-016-1299-1.

[14] M. Blanchin, A. Guilleux, J.-B. Hardouin, V. Sébille, Comparison of structural equation modelling, item response theory and Rasch measurement theory-based methods for response shift detection at item level: a simulation study, Stat. Methods Med. Res. 29 (4) (2020) 1015–1029, https://doi.org/10.1177/0962280219884574.

[15] M. Fokkema, N. Smits, H. Kelderman, P. Cuijpers, Response shifts in mental health interventions: an illustration of longitudinal measurement invariance, Psychol. Assess. 25 (2) (2013) 520–531, https://doi.org/10.1037/a0031669.

[16] B.L. King-Kallimanis, F.J. Oort, G.J.A. Garst, Using structural equation modelling to detect measurement bias and response shift in longitudinal data, AStA Adv. Stat. Anal. 94 (2) (2010) 139–156, https://doi.org/10.1007/s10182-010-0129-y.

[17] K. Hammas, V. Sébille, P. Brisson, J.-B. Hardouin, M. Blanchin, How to investigate the effects of groups on changes in longitudinal patient-reported outcomes and response shift using rasch models, Front. Psychol. 11 (2020) 3704, https://doi.org/10.3389/fpsyg.2020.613482.

[18] B.J. Taple, R. Chapman, B.D. Schalet, R. Brower, J.W. Griffith, The Impact of Education on Depression Assessment: Differential Item Functioning Analysis, Assessment. (2020) 1073191120971357. https://doi.org/10.1177/1073191120971357.

[19] K.J. Holzer, M.G. Vaughn, N.E. Fearn, T.M. Loux, M.A. Mancini, Age bias in the criteria for antisocial personality disorder, J. Psychiatr. Res. 137 (2021) 444–451, https://doi.org/10.1016/j.jpsychires.2021.03.025.

[20] A. Rouquette, J.-B. Hardouin, J. Coste, Differential item functioning (DIF) and subsequent bias in group comparisons using a composite measurement scale: a simulation study, J. Appl. Meas. 17 (2016) 312–334.
[21] G.H. Fischer, I. Ponocny, An extension of the partial credit model with an application to the measurement of change, Psychometrika 59 (2) (1994) 177–192, https://doi.org/10.1007/BF02295182.
[22] G.N. Masters, A rasch model for partial credit scoring, Psychometrika 47 (2) (1982) 149–174, https://doi.org/10.1007/BF02296272.
[23] M. Blanchin, J.-B. Hardouin, F. Guillemin, B. Falissard, V. Sébille, M. Gasparini, Power and sample size determination for the group comparison of patient-reported outcomes with rasch family models, PLoS ONE 8 (2) (2013) e57279, https://doi.org/10.1371/journal.pone.0057279.
[24] M. Blanchin, P. Brisson, ROSALI: Stata module to detect of response shift at item-level between two times of measurement, Statistical Software Components. (2020). https://ideas.repec.org/c/boc/bocode/s458796.html (accessed July 17, 2020).
[25] J.M. Bland, D.G. Altman, Multiple significance tests: the Bonferroni method, BMJ 310 (1995) 170, https://doi.org/10.1136/bmj.310.6973.170.
[26] R.D. Penfield, The impact of model misfit on partial credit model parameter estimates, J. Appl. Meas. 5 (2004) 115–128.
[27] M. Olsbjerg, K.B. Christensen, Modeling local dependence in longitudinal IRT models, Behav. Res. Methods 47 (4) (2015) 1413–1424, https://doi.org/10.3758/s13428-014-0553-0.
[28] C. Proust-Lima, V. Philipps, B. Perrot, M. Blanchin, V. Sébille, Continuous-time modeling of self-reported outcome data: a dynamic Item Response Theory model, ArXiv:2109.13064 [Stat]. (2021). http://arxiv.org/abs/2109.13064 (accessed October 20, 2021).